区間値距離を使用した階層的クラスタリングとその樹形図 (不確実性の下での意思決定の数理とその周辺)

AUTHOR(S):

小笠原, 悠; 久野, 優斗; 金, 正道

CITATION:

小笠原, 悠 ...[et al]. 区間値距離を使用した階層的クラスタリングとその樹形図 (不確実性の下での意思決定の数理とその周辺). 数理解析研究所講究録 2019, 2126: 19-27

ISSUE DATE:

2019-08

URL:

http://hdl.handle.net/2433/252228

数理解析研究所講究録 第2126巻 2019年 19-27

区間値距離を使用した階層的クラスタリングとその樹形図

(1) 首都大学東京 都市環境学部，(2) 弘前大学大学院理工学研究科
小笠原 悠 (1)，久野優斗 (2)，金正道 (2)
Yu Ogasawara(1), Yuto Hisano(2), Masamichi Kon(2)
(1) Faculty School of Urban Environmental Sciences, Tokyo Metropolitan University,
(2) Graduate School of Science and Technology, Hirosaki University

1 はじめに

クラスタリングは教師なし学習として有名な手法の一つである．その手法は現在まで様々なバリエーションが開発されており，機械学習分野で使用されている．クラスタリングは分割型と階層型に分かれるが，共に主に使用されるデータは実数が多い．[4] や [2] においてもデータのタイプは実数が中心に扱われる．一方で，実数のような伝統的なデータ以外にも，シンボリックデータを使用する方法も存在する．

シンボリックデータでは伝統的なデータタイプに対して，複数の値を持つ複合データや，データが区間を持つ区間値データが挙げられる [1]．区間値データは，データを区間で一様に捉えたものである．区間値データに対する統計手法は幾つか提案されている．その手法の中の一つとしてクラスタリングがある．区間値データに対するクラスタリングでは通常のクラスタリングと同様に分割型クラスタリングと階層型クラスタリングがある．

従来の区間値を使用したクラスタリングでは Gowda-Diday dissimilarity measure や Ichino-Yamaguchi dissimilarity measure，the generalized Minkowski distance, Housdorff distane が伝統的に使用される [1]．これらの距離・非類似度は入力は区間値をとるが，出力は実数値を取るものであった．本研究では入力と出力を共に区間値をとる区間値距離を提案する．更に，その区間値距離を利用した階層型クラスタリングとして，区間 Ward 法と，その樹形図作成方法を提案し，数値実験の結果を示す．

2 区間値の基本算法

最初に区間値の定義を示す．実数の集合を $\boldsymbol{R}$ とする．このとき区間値は以下で表される．

$$X = [\underline{X}, \overline{X}] = \{x \in \boldsymbol{R} : \underline{X} \leq x \leq \overline{X}\} \tag{2.1}$$

ただし，$a \in \boldsymbol{R}$ に対して，

$$a = [a, a]$$

とする．区間値の集合は

$$\boldsymbol{K} = \{[\underline{X}, \overline{X}] : \underline{X}, \overline{X} \in \boldsymbol{R}, \underline{X} \leq \overline{X}\}, \tag{2.2}$$

$$\boldsymbol{K}^p = \{[\underline{X}_1, \overline{X}_1], \cdots, [\underline{X}_p, \overline{X}_p] : \underline{X}_j, \overline{X}_j \in \boldsymbol{R}, \underline{X}_j \leq \overline{X}_j, j = 1, \cdots, p\} \tag{2.3}$$

で表される．以降は単純のため，$\boldsymbol{K} = \boldsymbol{K}^1$ とする．次に，区間値の基本算法 [3] を載せる．

定義 1. $* \in \{+, -, \times\}$ を実数の集合における二項演算子，加えて $X, Y \in \boldsymbol{K}, \lambda \in \boldsymbol{R}$ としたとき，

$$X * Y = \{x * y : x \in X, y \in Y\} \tag{2.4}$$
$$\lambda X = \{\lambda x : x \in X\}, \tag{2.5}$$
$$|X| = \{|x| : x \in X\} \tag{2.6}$$
$$\sqrt{x} = \{\sqrt{x} : x \in X\}, \underline{X} \geq 0 \tag{2.7}$$

とする．

$X, Y \in \boldsymbol{K}, \lambda \in \boldsymbol{R}$ に対して，$X = [\underline{X}, \overline{X}], Y = [\underline{Y}, \overline{X}]$ と置くと，定義 1 より

$$X + Y = [\underline{X} + \underline{Y}, \overline{X} + \overline{Y}] \tag{2.8}$$
$$\lambda X = \begin{cases} [\lambda \underline{X}, \lambda \overline{X}], & \lambda \geq 0 \\ [\lambda \overline{X}, \lambda \underline{X}], & \lambda < 0 \end{cases} \tag{2.9}$$
$$X - Y = [\underline{X} - \overline{Y}, \overline{X} - \underline{Y}] \tag{2.10}$$
$$|X| = \begin{cases} [\underline{X}, \overline{X}], & \underline{X}, \overline{X} \geq 0 \\ [0, \max\{-\underline{X}, \overline{X}\}], & \underline{X} < 0 < \overline{X} \\ [-\overline{X}, -\underline{X}], & \overline{X} \leq 0 \end{cases} \tag{2.11}$$
$$\sqrt{X} = [\sqrt{\underline{X}}, \sqrt{\overline{X}}], \underline{X} \geq 0 \tag{2.12}$$

が導かれる．上記から得られる算法を以下に定理として明示する．

定理 1. $X, Y, Z \in \boldsymbol{K}$ 及び $\lambda, \mu \in \boldsymbol{R}$ に対して，

$$X + Y = Y + X \tag{2.13}$$
$$(X + Y) + Z = X + (Y + Z) \tag{2.14}$$
$$\{0\} + X = X \tag{2.15}$$
$$\lambda, \mu \geq 0 \iff (\lambda + \mu)X = \lambda X + \mu X \tag{2.16}$$
$$\lambda(X + Y) = \lambda X + \lambda Y \tag{2.17}$$
$$(\lambda\mu)X = \lambda(\mu X) \tag{2.18}$$
$$1X = X \tag{2.19}$$
$$|X| = \{0\} \iff X = \{0\} \tag{2.20}$$
$$|\lambda X| = |\lambda||X| \tag{2.21}$$

が成り立つ．

本研究では区間値間の順序関係について $X = [\underline{X}, \overline{X}] \in \boldsymbol{K}, Y = [\underline{Y}, \overline{Y}] \in \boldsymbol{K}$ に対して

$$\underline{X} \leq \underline{Y}, \overline{X} \leq \overline{Y} \iff X \leq Y \tag{2.22}$$

とする．このとき $(\boldsymbol{K}, \leq)$ は半順序集合であることは明らかである．この順序関係に関する基本的な算法を定理 2 として提示する．

定理 2. $X, Y, Z, W \in \boldsymbol{K}, \lambda \in \boldsymbol{R}$ に対して

$$X \leq Y \to X + Z \leq Y + Z \tag{2.23}$$
$$X \leq Y, Z \leq W \to X + Z \leq Y + W \tag{2.24}$$
$$\lambda \geq 0, X \leq Y \to \lambda X \leq \lambda Y \tag{2.25}$$
$$\lambda \leq 0, X \leq Y \to \lambda X \geq \lambda Y \tag{2.26}$$
$$|X + Y| \leq |X| + |Y| \tag{2.27}$$

が成り立つ．

区間値の 2 乗の算法について述べる．2 乗の計算方法については以下の 2 つの方法が考えられる．

定義 2 (Moore et al.[3])**.** $X \in \boldsymbol{K}$ に対し，

$$X^2 = X \times X. \tag{2.28}$$

定義 3 (Moore et al.[3])**.** $X \in \boldsymbol{K}$ に対し，

$$X^{(2)} = \{x^2 : x \in X\}. \tag{2.29}$$

定義 2 は定義 1 を，そのまま適用したものであり，定義 3 は 2 乗を関数として捉えたものである．この 2 つの定義から得られる各値は等しくなるとは限らないことが既に知られている．例えば，$X = [-1, 1]$ の場合，$X^2 = [-1, 1] \neq X^{(2)} = [0, 1]$ となる．これらの 2 乗の定義と絶対値の定義から得られる特徴を以下に示す．

命題 1. $X \in \boldsymbol{K}$ に対して以下が成り立つ．

$$|X|^2 = |X|^{(2)}, |X|^2 = |X^2|, \tag{2.30}$$
$$|X|^{(2)} = |X^{(2)}| = X^{(2)}, \tag{2.31}$$
$$\sqrt{X^{(2)}} = |X|, \tag{2.32}$$
$$\sqrt{X^{(2)}} = \sqrt{X^2} = |X|, X \geq 0. \tag{2.33}$$

3 区間値距離

d を実数 $x \in \boldsymbol{R}$ の実数値関数とする．区間値 $x \in X, X \in \boldsymbol{K}$ の d における像は

$$d(x) = \{d(x) : x \in X\}$$

と表される [3]．この関数 d に距離関数を適用することで得られる入力と出力を共に区間値を取る区間距離関数を定義する．本研究では Ward 法を扱うため，d に平方ユークリッド距離を適用したものを使用する．

定義 4 (ユークリッド平方区間値距離 A, B)**.** $D_2^2 : \boldsymbol{K}^p \times \boldsymbol{K}^p \to \boldsymbol{K}, D_2^{(2)} : \boldsymbol{K}^p \times \boldsymbol{K}^p \to \boldsymbol{K}$ とし，$\boldsymbol{X} = (X_1, \cdots, X_p) \in \boldsymbol{K}, \boldsymbol{Y} = (Y_1, \cdots, Y_p) \in \boldsymbol{K}^p$ に対して

$$D_2^2(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{k=1}^{p} |X_k - Y_k|^2 \tag{3.1}$$

をユークリッド平方区間値距離 A,

$$D_2^{(2)}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{k=1}^{p} |X_k - Y_k|^{(2)} \tag{3.2}$$

をユークリッド平方区間値距離 B とする．

4　区間値統計量

本研究においては区間値を取る距離関数を扱うため，従来の区間値データの統計量とは異なる定義を使用する．

定義 5 (区間平均値)**.** $\boldsymbol{X} = (X_1, \cdots, X_p) \in \boldsymbol{K}^p$ に対して，

$$E(\boldsymbol{X}) = \frac{1}{p}\sum_{i=1}^{p} X_i \tag{4.1}$$

を $\boldsymbol{X}$ の区間平均値という．

定義 6. $\boldsymbol{X} = (X_1, \cdots, X_p) \in \boldsymbol{K}^p$ に対して，

$$V_1(\boldsymbol{X}) = \frac{1}{p}\sum |X_i - E(\boldsymbol{X})|^2, \tag{4.2}$$

$$V_2(\boldsymbol{X}) = \frac{1}{p}\sum |X_i - E(\boldsymbol{X})|^{(2)}, \tag{4.3}$$

$$V_3(\boldsymbol{X}) = \frac{1}{p}\sum (X_i - E(\boldsymbol{X}))^2, \tag{4.4}$$

$$V_4(\boldsymbol{X}) = \frac{1}{p}\sum (X_i - E(\boldsymbol{X}))^{(2)} \tag{4.5}$$

とし，それぞれを $\boldsymbol{X}$ の区間分散 *A,* 区間分散 *B,* 区間分散 *C,* 区間分散 *D* という．

区間分散 A, B は通常の分散の定義式を元にしていないことに注意する．

定義 7. $\boldsymbol{s} = (s_1, \cdots, s_p) \in \boldsymbol{K^p}$ に対して，$\sqrt{V(\boldsymbol{s})}$ を $\boldsymbol{s}$ の区間標準偏差と呼ぶ．

区間分散 A, B, C, D 間の関係については以下が成り立つ．

命題 2. $\boldsymbol{X} = (X_1, \cdots, X_p) \in \boldsymbol{K^p}$ に対して，$V_1(\boldsymbol{X}) = V_2(\boldsymbol{X}) = V_4(\boldsymbol{X})$ が成り立つ．

よって，命題 1, 2 より

$$\begin{aligned} D_2^{(2)}(\boldsymbol{X}, \boldsymbol{Y}) &= \sum_{k=1}^{p} |X_k - Y_k|^{(2)} \\ &= \sum_{k=1}^{p} (X_k - Y_k)^{(2)} \end{aligned}$$

が成り立つことから，定義 3 を使用することにより，平方ユークリッド区間値距離は通常の偏差平方和を区間値で考えたものと一致することがわかる．以降は平方ユークリッド区間値距離 B を使用する．

5　区間 Ward 法

本研究では区間値の順序関係は半順序のみを扱うため，以下に示す極小を使用する．

定義 8. $\chi \subset \boldsymbol{K}, X \in \chi$ であるとき，X が χ の極小要素であるとは，

$$Y \leq X, Y \in \chi \to Y = X$$

であるときを言う．χ の極小要素全ての集合を Min χ と表す．

クラスターの集合を $\mathcal{G} = \{G_1, \cdots, G_k\}$ と表記する．このとき個体は $\alpha \in \boldsymbol{K}^p, \alpha \in G_i, i = 1, \cdots, k$ とする．クラスター G_g に属する個体数を n_g とすると，全体の個体数は $n = \sum_{g=1}^{k} n_g$ となる．本研究では区間値を用いた Ward 法を 2 種類（区間 Ward 法-A，区間 Ward 法-B）提示する．最初に，区間値による偏差平方和を使用する区間 Ward 法-A を示す．

偏差平方和を

$$ESS(G_g) = \sum_{\alpha \in G_g} D_2^{(2)}(\alpha, E(G_g))$$

とする．ここで，$E(G_g) = (E(\alpha_{1,1}, \cdots, \alpha_{n_g,1}), \cdots, E(\alpha_{1,p}, \cdots, \alpha_{n_g,p})), \alpha_i = (\alpha_{i,1}, \cdots, \alpha_{i,p}), i = 1, \cdots, n_g$ とする．クラスター $G_f \in \mathcal{G}$ と $G_g \in \mathcal{G}$ の間の非類似度を以下で定義する．

$$D_{2a}(G_f, G_g) = ESS(G_f \cup G_g) - ESS(G_f) - ESS(G_g) \tag{5.1}$$

この，クラスター間の非類似度として (5.1) を使用する方法を区間 Ward 法-A と呼ぶ．区間 Ward 法-A は通常の偏差平方和の式を利用した $ESS(\cdot)$ を使用しておりデータが間隔尺度の場合は基本算法の延長となることに注意する．よって，区間 Ward 法-A は通常の Ward 法と同様の解釈で計算出来る．しかし，非類似度の計算において区間値の減算が行われているため非類似度の下限は 0 以下の値を取ることに注意する．

非類似度の下限が必ず 0 以上の値を取る方法として，以下の区間 Ward 法-B を合わせて示す．$\boldsymbol{x} = (x_1, \cdots, x_m)^T, \boldsymbol{y} = (y_1, \cdots, y_n)^T, x_i, y_j \in R^p, i = 1, \cdots, m, j = 1, \cdots, n$ とし，

$$\begin{aligned} f_{mn}(\boldsymbol{x}, \boldsymbol{y}) = &\sum_{j=1}^{p}\sum_{i=1}^{m}(x_{ij} - \mu_{XYj})^2 + \sum_{j=1}^{p}\sum_{i=1}^{n}(y_{ij} - \mu_{XYj})^2 \\ &- \sum_{j=1}^{p}\sum_{i=1}^{m}(x_{ij} - \mu_{Xj})^2 - \sum_{j=1}^{p}\sum_{i=1}^{n}(y_{ij} - \mu_{Yj})^2 \end{aligned} \tag{5.2}$$

とする．$G_f = \{X_1, \cdots, X_m\}, G_g = \{Y_1, \cdots, Y_m\}$, $X_i, Y_j \in \boldsymbol{K}^p, i = 1, \cdots, m, j = 1, \cdots, n$ が与えられたとき，

$$\Delta(G_f, G_g) \tag{5.3}$$

$$= \{f_{mn}(\boldsymbol{x}, \boldsymbol{y}); x_i \in X_i, i = 1, \cdots, m, y_j \in Y_j, j = 1, \cdots, n\} \tag{5.4}$$

とする．$\Delta(G_f, G_g)$ が取りうる範囲を非類似度

$$D_{2b}(G_f, G_g) = [\min \Delta(G_f, G_g), \max \Delta(G_f, G_g)] \tag{5.5}$$

と定義する．非類似度 $D_{2b}(G_f, G_g)$ を使用した手法を区間 Ward-B と呼ぶ．この非類似度は

$$\min\ f_{mn}(\boldsymbol{x}, \boldsymbol{y}) \tag{5.6}$$

$$s.t.\ x_i \in X_i, y_j \in Y_j, i = 1, \cdots, m, j = 1, \cdots, n \tag{5.7}$$

と

$$\max\ f_{mn}(\boldsymbol{x}, \boldsymbol{y}) \tag{5.8}$$

$$s.t.\ x_i \in X_i, y_j \in Y_j, i = 1, \cdots, m, j = 1, \cdots, n \tag{5.9}$$

を解くことで得られる．$f_{mn}(\boldsymbol{x},\boldsymbol{y})$ は [2] より平均値ベクトル μ_{G_f},μ_{G_g} による式に変形できることから，問題 (5.6) と (5.8) はそれぞれ，

$$\min \ \frac{mn}{m+n}(\mu_{G_f}-\mu_{G_g})(\mu_{G_f}-\mu_{G_g})^T \tag{5.10}$$

$$s.t.\ \mu_{G_f}\in E(G_f),\mu_{G_g}\in E(G_g) \tag{5.11}$$

と

$$\max \ \frac{mn}{m+n}(\mu_{G_f}-\mu_{G_g})(\mu_{G_f}-\mu_{G_g})^T \tag{5.12}$$

$$s.t.\ \mu_{G_f}\in E(G_f),\mu_{G_g}\in E(G_g) \tag{5.13}$$

になり，決定変数の数は個体数に依存しない．(5.10) と (5.12) は半正定値問題であることに注意する．

これらの区間 Ward 法-A と区間 Ward 法-B を使用したアルゴリズムを具体的に示す．単純のため，$D\in\{D_{2a},D_{2b}\}$ とする．

S1 初期設定として個々の個体をクラスターとする．すなわち，

$$G_i:=\{i\},i\in N$$

とし，$k:=n$，つまり，クラスターの数は個体数とする．

S2 非類似度極小のクラスター対を探して結合する．つまり，

$$\chi:=\{D(G_i,G_{i'}):G_i,G_{i'}\in\mathcal{G},G_i\neq G_{i'}\}$$

とし，$D(G_g,G_r)\in \mathrm{Min}\ \chi$ となる G_g と G_r を $\mathcal{G}$ から取り除き，$G'=G_g\cup G_r$ を $\mathcal{G}$ に追加する．ここで $k:=k-1$ とし，クラスターを 1 つ減らす．このとき $k=1$ ならば終了する．このときの結合するときの区間値距離 $D(G_g,G_r)$ を結合レベルと呼ぶ．

S3 全ての $G_i\in\mathcal{G},G_i\neq G'$ についてクラスター間の非類似度を再計算し，S2 に戻る．

ステップ S2 では極小を利用しているため，結合候補となるペアが複数存在することがある．本研究ではその候補の中から結合するペアを決定する方法を以下に 4 つ設定した．

1. Min χ が与えられたとき，$D(G_g,G_r)\in \underset{X\in\mathrm{Min}\ \chi}{\arg\min}\ \min X$ を満たす G_g と G_r を選ぶ．
2. Min χ が与えられたとき，$D(G_g,G_r)\in \underset{X\in\mathrm{Min}\ \chi}{\arg\min}\ \max X$ を満たす G_g と G_r を選ぶ．
3. Min χ が与えられたとき，$D(G_g,G_r)\in \underset{X\in\mathrm{Min}\ \chi}{\arg\min}\ m(X)$ を満たす G_g と G_r を選ぶ．ここで $X=(\underline{X},\overline{X})\in\boldsymbol{K}$ に対して $m(X)=\frac{1}{2}(\underline{X}+\overline{X})$ とする．
4. Min χ の中からランダムに一つの元を選ぶ．

これらの設定を本研究ではそれぞれ，minmin, maxmin, midmin, random と呼ぶこととする．

6 樹形図作成方法

通常のデンドログラムのようなノードが点のみを示す方法では，区間値距離を用いた結果を描写することは出来ない．よって本研究では幅を持つ結合レベルを表現するために，結合レベルの描写を変更した新たなデンドログラムの描写方法を提案する．デンドログラムの結合レベルの描写について，以下のルールを適用する．

1. 結合レベルの幅は上限と下限まで伸ばした両矢印で表現する．両矢印を描写するクラスターは結合するクラスターの片側のみとし，全ての結合について両矢印を描写する側は統一することとする．更に，ノードは全て結合レベルの上限部分で描写することとする．このルールに従った例が以下の図 1 である．

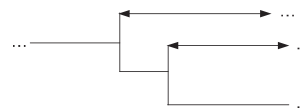

図 1: ルール 1 の適用例

2. ルール 1 においてはデンドログラムを作成する際には図 1 と同様に上側のクラスターを矢印にするように統一しているものとする．ここで，あるステップ 2 を考える．その時の結合するクラスターを G_g と G_r とし，レベルを $D(G_g, G_r)$ とする．デンドログラムを作成する際には上側と下側，それぞれどちらを G_g もしくは G_r とするのかは任意ではあるが，ここでは G_g を上側，すなわち矢印で表記するものとする．$n_g = 1$ である場合にはデンドログラム作成の際，$D(G_g, G_r)$ の上限と下限の範囲で矢印を引く．このときの例を図 2 に示す．

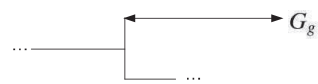

図 2: $n_g = 1$ の例

次に $n_g > 1$ の場合を考える．このとき，$G_g = G_k \cup G_l$ とする．$D(G_g, G_r)$ の上限と下限をそれぞれ $\overline{D(G_g, G_r)}$ と $\underline{D(G_g, G_r)}$ とし，$D(G_k, G_l)$ の上限と下限をそれぞれ $\overline{D(G_k, G_l)}$ と $\underline{D(G_k, G_l)}$ とする．ここで，$\underline{D(G_g, G_r)} \leq \overline{D(G_k, G_l)}$ のときは，$\underline{D(G_g, G_r)}$ から $\overline{D(G_k, G_l)}$ のノードまでの間を破線で繋ぐこととする．このケースを図 3 で示す．$\underline{D(G_g, G_r)} > \overline{D(G_k, G_l)}$

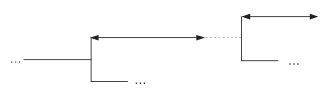

図 3: $n_g > 1$ かつ $\underline{D(G_g, G_r)} \leq \overline{D(G_k, G_l)}$ の例

の場合は，$\overline{D(G_g, G_r)}$ から $\overline{D(G_k, G_l)}$ までの矢印を実線で表し，$\overline{D(G_k, G_l)}$ から $\underline{D(G_g, G_r)}$ は矢印を破線で表すことにする．このケースを図 4 に示す．

これらのルールを各結合ノードにて再帰的に実行することによって，デンドログラムを作成する．当然，結合する際に連続して結合レベルの差が小さい場合はこれらのルールでは線が重なる場合

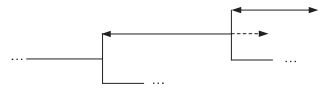

図 4: $n_g > 1$ かつ $\underline{D(G_g, G_r)} > \overline{D(G_k, G_l)}$ の例

が生じることに注意する．これらのルールにより結合レベルが区間値になっていても従来のデンドログラムと同様の直感的な理解が可能となる．

7　数値実験

小規模な数値実験により，区間 Ward 法-A と区間 Ward 法-B により得られるデンドログラムを提示する．入力データは $X_1, X_2 \in \boldsymbol{K}$ とし，表 1 の値をとるものとする．このデータに区間 Ward 法-A と区間 Ward 法-B を適用し，本研究で提案したデンドログラム作図法により図示したものが図 5 と図 6 である．極小を選択する設定は minmin を使用した． 図 5 の結合レベル零

表 1: 入力データ X_1, X_2

	$\underline{X}_1$	$\overline{X}_1$	$\underline{X}_2$	$\overline{X}_2$
A	14.9	20.1	2.9	5.6
B	12.6	19.7	3.5	6.1
C	20.7	24.9	1.9	5.4
D	10.7	19.1	2.4	6
E	12.4	19.6	2.7	5.6
F	21.7	24	2.7	5.6
G	18.1	21	6.3	7.1
H	19	20.5	5.9	6.8

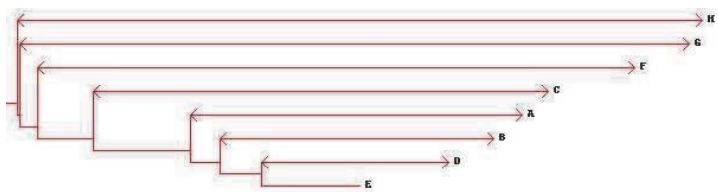

図 5: 区間 Ward-A のデンドログラム

は個体 E のラベルの位置である．この事実から，結合レベルの下限が負に割り込んでいることを見ることが出来る．一方，区間 Ward 法-B は概ね直観に沿う結果となっており，見た目も通常の Ward 法の結果に似ていることが分かる．図 5 と図 6 で結合している順番を比較すると，この二つの手法は大きく異なっている．実際にこの 2 つの手法間の Fowlkes-Mallows index (FMI) は $FMI_i = (0.600, 0.639, 0.424, 0.235, 0.000, 0.000), i = 2, \cdots, 7$ となっており，最初のステップから異なる結合ペアが選ばれていることが分かる．ここでの i はクラスター数を表す．次に表 1 の各項目の上限下限に対して，平均 5, 分散 3 の正規分布に従う乱数を加えたデータを用いて計算を行い．そこから得られる FMI を各階層別の箱ひげ図として示す．図 7 は設定 minmin を使用したもので

図 6: 区間 Ward-B のデンドログラム

あり，横軸はクラスター数，縦軸は FMI である．図 7 を見ると，最初のループから類似度の分布

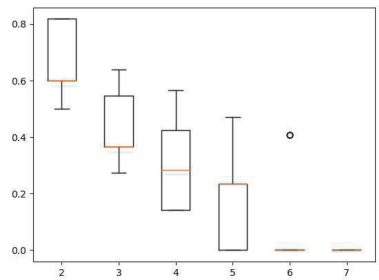

図 7: 区間 Ward 法-A と区間 Ward 法-B の minmin による FMI

は総じて低いことがわかる．設定 1 はループの最初は類似度は低いものの，結合が進むにつれて上がっていき，おおよそ最後の結合前で類似性が高くなる．

8 まとめ

本研究では入力と出力が共に区間値を取る区間値距離として，平方ユークリッド区間値距離を提案した．それに伴い，新たな統計量として区間平均値と区間分散を定義し，その特徴を示した．更に，平方ユークリッド区間値距離を利用した Ward 法を元に，区間 Ward 法-A と区間 Ward 法-B の 2 つのクラスタリング方法を提案し，それらを使用した小規模な数値実験を行った．今後の課題として，区間値解析で既に得られている特徴や従来の Ward 法との関係性を明らかにすることや，区間 Ward 法の単調性の有無を明らかにすることなどが考えられる．

参考文献

[1] L Billard and E Diday. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, 2006.

[2] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*, volume 20. Siam, 2007.

[3] Ramon E Moore, R Baker Kearfott, and Michael J Cloud. *Introduction to interval analysis*, volume 110. Siam, 2009.

[4] Konstantinos Koutroumbas Sergios Theodoridis. *Pattern Recognition, Fourth Edition*. Academic Press, 2008.